HHS Public Access

Author manuscript

Curr Protoc Hum Genet. Author manuscript; available in PMC 2018 June 19.

Published in final edited form as:

Curr Protoc Hum Genet. ; 95: 1.22.1–1.22.23. doi:10.1002/cphg.48.

Population Stratification in Genetic Association Studies

Jacklyn Hellwege[1], **Jacob Keaton**[1], **Ayush Giri**[1], **Xiaoyi Gao**[2], **Digna R. Velez Edwards**[3], and **Todd L. Edwards**[1,†]

[1]Vanderbilt Genetics Institute, Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37203, USA

[2]Department of Ophthalmology and Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA

[3]Vanderbilt Genetics Institute, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN 37203, USA

Abstract

Population stratification (PS) is a primary consideration in studies of the genetic determinants of human traits. Failure to control for it may lead to confounding, causing a study to fail for lack of significant results or resources to be wasted following false positive signals. Here we review historical and current approaches for addressing PS when performing genetic association studies in human populations. We describe methods for detecting the presence of PS including global and local ancestry methods. We also describe approaches for accounting for PS when calculating association statistics, such that measures of association are not confounded. Many traits are being examined for the first time in minority populations, populations that may inherently feature PS.

Keywords

POPULATION STRATIFICATION; ASSOCIATION CONFOUNDING; GLOBAL ANCESTRY; LOCAL ANCESTRY; ADMIXTURE; ADMIXTURE MAPPING

KEY CONCEPTS

Definition and Causes of Population Stratification

As *Homo sapiens* geographic range expanded over time and groups left the site of their geographic origins in Africa [Vigilant, et al. 1991], they separated into subgroups and experienced novel stresses and environments. Geographic isolation, interbreeding, and adaptation differentiated human populations from each other [Schlebusch, et al. 2012].

Fossil and genetic evidence suggests that anatomically modern humans evolved in Africa about 150,000 to 190,000 years ago [McDougall, et al. 2005; White, et al. 2003] and expanded into a diverse array of niches there, providing Africans with the highest level of

†Correspondence should be addressed to: Todd L. Edwards, PhD, Assistant Professor, Division of Epidemiology, Department of Medicine, Vanderbilt Genetics Institute, Institute for Medicine and Public Health, Vanderbilt University, 2525 West End Ave., Suite 600 6th floor, Nashville, TN 37203, P 615.322.3652, todd.l.edwards@vanderbilt.edu.

genetic diversity among current human continental populations [Rosenberg, et al. 2002; Tishkoff, et al. 2009]. Humans subsequently migrated into Europe, Asia, and the Americas in an approximately West-to-East pattern that began approximately 50,000 – 100,000 years ago [Gravel, et al. 2011; Harris and Nielsen 2013; Li and Durbin 2011; Mallick, et al. 2016] and concluded with the settlement of South America sometime in the last 15,000 years [Jenkins, et al. 2012]. The features of this scenario are increasingly complex, as understanding of hominin origins are updated regularly by increasingly sophisticated studies of modern populations, discoveries of ancient DNA specimens, and archaeological artifacts. A recent review by Nielsen et al. covers this history and the evidence that supports it [Nielsen, et al. 2017].

Among the effects of this period of colonization and the migrations during and afterward, as well as mating between populations of humans and other hominins [Green, et al. 2010; Meyer, et al. 2012; Vernot and Akey 2015], are differences across populations in allele frequencies throughout the genome. These differences, however they arise, are detectable in studies of human populations and provide information about both demographic history and geographic origins in modern humans [Novembre, et al. 2008; Wang, et al. 2012a]. This state, where populations are distinguishable by observing genotypes, is referred to as population structure or population stratification (PS).

PS may confound associations between genotype and the trait of interest in a genetic study. When PS exists, false positive or negative associations between genotype and trait may arise from differences in local ancestry that are unrelated to disease risk or trait variance. A consequence of PS, genetic admixture, arises from interbreeding of ancestral groups. A common example of genetic admixture is the African American population, which has both African and European ancestry. These factors must be considered in study designs and accounted for statistically in order for results of genetic association studies to be reliable. In this Unit, we will discuss the causes of PS and its history in genomic investigations, methods for observing global and local ancestry within a population, and techniques to account for and leverage differences in ancestry within genetic association studies.

PS is caused by non-random mating and most often arises due to geographic isolation of subpopulations with low rates of migration and gene flow over the course of several generations (Hartl and Clark, 2007). The geographic separation of these isolates allows for divergent random genetic drift due to sampling error in the set of parental alleles, which is subsequently propagated through successive generations. As a result, allele frequencies change randomly over time as an independent process for each population isolate, ultimately causing observable differences in the frequency of many alleles after several generations of separation and differentiation.

This scenario also introduces the possibility of selection for different traits in different geographic regions. A classic example of selection is hypolactasia, or lactase intolerance, a trait which prevents individuals from metabolizing the milk sugar lactose into adulthood through decreased production of the lactase enzyme [Bayless, et al. 2017]. One of the first genetic variants found to be associated with hypolactasia in humans, rs4988235, resides not in the lactase gene *LCT*, but rather in an enhancer region within an intron of another gene,

MCM6, approximately 15kB upstream of the *LCT* promoter [Enattah, et al. 2002; Lewinsky, et al. 2005; Olds and Sibley 2003]. A two-variant haplotype including rs4988235 and rs182549 explains 77% of hypolactasia variance in Europeans, but does not explain the trait distribution in individuals of African ancestry [Mulcare, et al. 2004]. Multiple studies have discovered additional variants including rs145946881, rs41380347, and rs41525747 that explain the distribution of hypolactasia in Africans, all of which reside in the same enhancer region as the European variant rs4988235 [Friedrich, et al. 2012; Ingram, et al. 2007; Ingram, et al. 2009; Tishkoff, et al. 2007]. Age estimates for the European hypolactasia variant rs4988235 range from 2,188 to 20,650 years ago [Bersaglieri, et al. 2004]. Similarly, age estimates for the African variant rs145946881 range from 1,200 to 23,200 years ago [Tishkoff, et al. 2007]. An empirical example of PS is the spurious association between *LCT* and height in a case-control study of European American population [Campbell, et al. 2005]. A single nucleotide polymorphism (SNP) in *LCT* showed strong association (p-value < 10^{-6}) with height without addressing PS. No significant association was detected between the SNP and height after correcting for PS.

Larger, more ancient gene pools, such as African ancestry, have a greater amount of overall variation and a finer linkage disequilibrium (LD) structure between markers [Goddard, et al. 2000]. Maximum ability to differentiate populations comes from genetic markers with large frequency differences among the parental populations for admixed samples. These markers, often SNPs, are known as ancestry informative markers (AIMs). AIMs are frequently incorporated into genotyping experiments when PS is suspected for downstream conditioning on inferred ancestral information in association modeling [Pritchard and Donnelly 2001].

The differentiation among subpopulations is detectable even when the regional differences are subtle, as has been described in Chinese and Japanese and European populations [Gao and Starmer 2007]. Cultural differences among populations also create stratification, even when populations inhabit the same geographical region. An example of this is the detectable differences among populations that speak Khoesan languages that include click-consonants from non-Khoesan speaking peoples who occupy the same geographic range [Tishkoff, et al. 2009]. Recent evidence shows that, even after correction for ancestry inferred from common genetic factors such as AIMs, subtle uncorrected population substructure persists in some genomic studies [Bhatia 2016].

Measures of genetic differentiation

There are several measures of genetic differentiation to evaluate the relationship of subpopulations to one another. One of the classical approaches is the fixation index (F_{st}), which compares the differences in expected heterozygosity across populations under Hardy-Weinberg Equilibrium [Weir and Cockerham 1984; Weir and Hill 2002; Wright 1921]. The drift toward fixation in isolated groups results in a loss of heterozygosity in the total population, which is known as the Wahlund effect [Wahlund 1928]. Specifically, F_{st} quantifies the proportional impact the subpopulations have on the heterozygosity estimate relative to the situation where there was no population structure. An expression for F_{st} relating the expected heterozygosity under Hardy-Weinberg Equilibrium H of a single

marker in the subpopulation s, denoted H_s, to the total H_t is $F_{st} = H_t - H_s / H_t$. Average F_{st} across a set of unlinked markers is a standard metric for assessing population genetic differentiation. Smaller F_{st} indicates similar allele frequencies between populations, while larger values mean that the allele frequencies are different [Holsinger and Weir 2009]. Sewell Wright suggested the following guidelines for interpreting values of F_{st}: 0–0.05 indicates little differentiation, 0.05–0.15 indicates moderate differentiation, 0.15–0.25 indicates great differentiation, and greater than 0.25 indicates very great differentiation.

Because the effects of alleles on traits detected in genetic studies are usually subtle, relatively small levels of differentiation can confound tests of association. Factors that can accelerate the rate of differentiation at a locus are small subpopulation size, inbreeding, selection, and mutation. Some factors that slow the rate of differentiation are migration and gene flow between subpopulations and large population size. Approaches for using F_{st} for estimating migration rates, inferring demographic history, identifying genomic regions under selection, forensic science and association mapping, and a discussion of the relationship with coalescent theory were reviewed by Holsinger and Weir [Holsinger and Weir 2009]. Further, observed F_{st} across human subpopulations have also been reported [Steele, et al. 2014].

Another quantification of the differences between population samples is the allele sharing distance (ASD) [Gao and Martin 2009; Gao and Starmer 2007]. ASD is a pair-wise measure among subjects across a large set of markers, and is defined by the expression, where $d_l = 0$ if two individuals have two alleles in common at the l-th locus; $d_l = 1$ with one allele in common, and $d_l = 2$ when there are no alleles in common. The relationship between ASD and the closely related identical by state (IBS) has been described by Miclaus et al [Miclaus, et al. 2009].

Admixture and Admixture Mapping

Although it simplifies the description of PS to imagine the allele frequencies of distinct subpopulations randomly drifting away from each other over time, populations also tend to mix. This is known as admixture, and at the first generation after two distinct populations begin mixing, these offspring have half of their genetic material from each of the maternal and paternal populations. In subsequent generations, average ancestral proportions in offspring vary according to the composition and rates of genetic exchange among the ancestral populations.

African Americans are a classic example of this, where approximately 80% of the genome is derived from African ancestors and 20% from European ancestors at autosomes, and there are greater proportions of African-derived X chromosomes due to historically skewed transmission to offspring from European males and African females [Bryc, et al. 2010].

Examples of Population Stratification in Genetic Studies

As a simple numeric example of PS, suppose some data are collected as listed in Table 1. In population 1, the cell frequency (case, allele A) is 0.27, which is equal to the product of the marginal frequencies 0.3*0.9. This relationship holds for population 2, i.e. 0.08 = 0.8*0.1. Therefore, no association exists between marker alleles and case-control status. However, in

the pooled data of population 1 and 2, the cell frequency for (case, allele A), 0.175, is no longer equal to the product of the marginal frequencies 0.55*0.5 and a chi-square test with one degree of freedom is significantly association with p-value < 0.0001. Therefore, even though there is no association in either population 1 or 2, a false positive association exists in the pooled population.

Confounding due to PS resulting in spurious genotype-phenotype associations is well-documented. A classic example is a study by Knowler *et al.* that describes an association between a polymorphism in the immunoglobulin Gm system, $Gm^{3:5,13,14}$, and type 2 diabetes in Native Americans recruited from the Gila River Indian Community in southern Arizona [Knowler, et al. 1988]. Gm polymorphisms have different frequencies between ancestry groups [Brucato, et al. 2009; Schanfield and Kirk 1981; Williams, et al. 1985]. Knowler *et al.* showed that $Gm^{3:5,13,14}$ was not a causal genetic factor in the development of type 2 diabetes, but that the observed association was confounded by admixture between Native American and European ancestry groups [Knowler, et al. 1988]. After adjustment for admixture proportions, the association was no longer statistically significant.

The spurious association of markers that are highly variable between ancestry groups is not uncommon. Choudry *et al.* analyzed AIMs for association with asthma in two admixed Latino populations, Mexicans and Puerto Ricans, which have the highest and lowest asthma morbidity, mortality, and prevalence rates among all US populations, respectively [Choudhry, et al. 2006; Homa, et al. 2000; Moreno-Estrada, et al. 2013]. Of all 44 AIMs tested, eight were significantly associated with asthma, but only two remained significant after adjustment for PS.

Some populations have very complex recent demographic histories that must be accounted for in statistical analyses. For example, the Brazilian population is made up of individuals with varying proportions of African, Native American, and European ancestry [Pena, et al. 2011]. Skin color is poorly correlated with genetic ancestry in the Brazilian population and therefore self-reported race can be inaccurate for genetic studies [Pena, et al. 2011]. Early genetic studies of type 1 diabetes (T1D) in Brazilians reported geographic variability in HLA-DR and HLA-DQ allele frequencies, two genetic loci strongly associated with T1D in Europeans [Silva, et al. 2008; Thomson, et al. 2007]. In a study accounting for PS, Gomes *et al.* identified a novel protective haplotype DRB1*10-DQB1*0501 [Gomes, et al. 2017].

Accounting for PS in candidate gene studies is challenging due to the lack of genome-wide coverage of genetic factors from which ancestry may be inferred. A classic example is the observed association between a restriction fragment length polymorphism (RFLP) upstream of the insulin gene *INS* and T1D [Bell, et al. 1984]. Replication of this association was consistently reported in population-based studies across several ancestries, but no evidence of linkage was detected in family studies [Spielman, et al. 1989]. These findings initially suggested that the observed association was the result of confounding due to PS. However, implementation of the transmission disequilibrium test (TDT), a linkage method that incorporates family member controls and is robust top PS, detected strong evidence of linkage between the RFLP and T1D [Spielman, et al. 1993]. The failure of previous family-based genetic studies to detect linkage between the RFLP and T1D was likely due to a lack

of power to detect variants with modest effects. Recent studies have shown that as few as 30 AIMs are sufficient to accurately estimate ancestry proportions in African American populations, suggesting that modest numbers of AIMs are adequate in more complex populations [Kodaman, et al. 2013; Ruiz-Narvaez, et al. 2011].

Patterns of PS may also provide insights into demographic histories in admixed populations. An analysis of 128 AIMs in the Cuban population showed a large European paternal contribution and a large Native American and African maternal contribution [Marcheco-Teruel, et al. 2014]. These contributions are concordant with the historical context of male European settlers mating with Native American females during the early stages of colonization, and later mating with African females during the period of transatlantic slave trade [Benn-Torres, et al. 2008; Mendizabal, et al. 2008]. Similarly, analysis of genetic data from 23 and Me (23 and Me Inc., Mt. View, CA), a direct-to-consumer genetic testing company, shows evidence of sex-biased gene-flow in the U.S. reflective of early colonization by and subsequent immigration of European populations [Bryc, et al. 2015; Eriksson, et al. 2010; Tung, et al. 2011].

QUANTIFYING POPULATION STRATIFICATION

Global and Local Ancestry

Global ancestry—Many methods for working with PS estimate global parameters to summarize the ancestry of study subjects. These parameters are often useful for both PS detection and statistical control of confounding by PS. Depending on the questions addressed; methods to detect and quantify PS require genotype data from a handful of carefully selected genetic variants to a large number of genome-wide SNPs. A common question in PS detection is the number and type of SNPs needed to detect PS in a given context. Regardless of the statistical methods employed, the more similar two populations are, the more markers need to be evaluated to detect the differences.

If the study is performing small-scale genotyping, then AIMs may be the most cost-efficient way to quantify ancestry. This approach is only possible if AIMs have been identified *a priori*, as has been done for the reference populations from the International HapMap Project (http://www.hapmap.org). The 1000 Genomes Project [Genomes Project, et al. 2012] is also widely used for most populations, with the Haplotype Reference Consortium (HRC) panel [McCarthy, et al. 2016] becoming more commonly utilized recently. However, construction of population-specific reference sets through whole genome sequencing is becoming increasingly more common [Low-Kam, et al. 2016] (French-Canadian); [Tang, et al. 2016] (Australian Aboriginal; exome); [Higasa, et al. 2016] (Japanese); [Thareja, et al. 2015] (Persian Kuwaiti); [Huang, et al. 2015] (UK10K, United Kingdom), [Kawai, et al. 2015] (1KJPN, Japanese); [Wong, et al. 2014] (South Asian Indians); [Kim, et al. 2014] (Korean); [Deelen, et al. 2014] (GoNL, Netherlands); [Carmi, et al. 2014] (Ashkenazi); [Wong, et al. 2013] (Asian Malays)), though some of these (i.e. UK10K and GoNL) have also been included in the HRC. Otherwise if genome-wide association study (GWAS) data are available, then 50,000 to 100,000 linkage disequilibrium (LD)-pruned SNPs may be used to estimate global ancestry.

Local ancestry—With the availability of GWAS data in admixed populations and advances in admixture mapping, several methods have been developed to classify ancestry in small chromosomal regions. Early methods evaluating local ancestry in admixed populations, including MALDsoft, STRUCTURE, and ANCESTRYMAP, were based on Hidden Markov Models (HMM) [Falush, et al. 2003; Hoggart, et al. 2004; Montana and Hoggart 2007; Patterson, et al. 2004; Zhu, et al. 2006].

Local ancestry estimates can be used as covariates in linear models on a SNP-by-SNP basis [Wang, et al. 2011]. Alternately, tests based on a conditional likelihood framework, which models the distribution of the test SNP given disease status and flanking marker genotypes, are also available [Wang, et al. 2011]. Alternatively, principal components analysis (PCA), multidimensional scaling (MDS), STRUCTURE, and other methods can provide estimates of global ancestry in that are useful for adjusting for PS in linear models. Local ancestry estimates the ancestral origin of chromosomes at a locus. While adjusting for global estimates is the most common approach and controls confounding in GWAS, residual confounding might lead to increased type II errors, and improvements in power have been noted for adjusting for local estimates [Wang, et al. 2011].

Global Ancestry Methods

Methods for estimating ancestry proportions—Direct evaluation of genetic ancestry proportions involves comparisons of sample data to reference allele frequencies are based on the use of AIMs. The necessary difference in allele frequency to differentiate two populations can vary depending on the number of AIMs and the genetic distance between populations. Often a 20% difference in allele frequencies is used to define AIMs. AIMs can be identified from published lists, or through empirical assessment of allele frequencies in the available genetic data. Fewer AIMs are required to quantify global ancestry when allele frequency differences are large. However, inclusion of more AIMs increases the precision of ancestry estimates. Accurate estimation of ancestry proportions is also dependent on the number of parental populations (designated K) assumed to contribute to the overall genetic ancestry of the target population. Most of the approaches described here can be implemented either defining K to the number of suspected subpopulations believed to be present in the data and providing those K reference datasets, or alternatively, selecting increasing values for K, and choosing the value of K where the likelihood of the data given K is largest. If the likelihood is largest at K=1, then there is no evidence for PS. Subjects who map into the known groups can then be identified as a member of that population or a related subpopulation. Several software packages exist for computation of global genetic ancestry proportion estimates, among them the most popularly adopted are STRUCTURE [Falush, et al. 2003; Pritchard, et al. 2000a], FRAPPE [Tang, et al. 2006] and ADMIXTURE [Alexander and Lange 2011; Alexander, et al. 2009].

STRUCTURE uses a Bayesian approach and relies on a Markov Chain Monte Carlo (MCMC) algorithm to jointly sample the posterior distribution of allele frequencies and fractional group memberships. STRUCTURE assumes that the data are comprised of mosaics of chromosomes from an arbitrary number of homogeneous ancestral subpopulations (K), and that each subpopulation is characterized by a distinct vector of

allele frequencies. STRUCTURE is sensitive to non-random missing data, and running the software with enough iterations to ensure the convergence of the MCMC algorithm is also of concern when utilizing this method [Yang, et al. 2005]. In practice, between 1,000 and 5,000 iterations are sufficient for burn-in, and 5,000–10,000 are sufficient for estimation. Too little iteration will cost accuracy, while too much iteration will only cost computational run-time, so if in doubt, use of more iterations than necessary will yield accurate results.

One of the earliest approaches for controlling for global ancestry was Structured Association (SA) and is a two-step procedure [Pritchard and Donnelly 2001; Pritchard, et al. 2000b]. The first step uses markers that are not associated with any trait of interest to assign individuals to subpopulations and then test for association within those groups. The first step can be performed using the program STRUCTURE. The second step, the association testing, is performed using likelihood ratio tests. The most challenging issue in SA analyses is to estimate the correct number of subpopulations to condition on. STRUCTURE does provide a data-driven way to infer this number, by scanning over choices of K, and choosing the value that maximizes the likelihood of the data given K.

FRAPPE and ADMIXTURE each use a maximum likelihood estimate (MLE) approach and optimize the likelihood for both allele frequencies and fractional group memberships using an expectation-maximization (EM) algorithm, but ADMIXTURE uses a faster optimization algorithm. ADMIXTURE yields ancestry estimates with similar accuracy as STRUCTURE but uses less computing time, and has many of the same capabilities, including the ability to estimate the number of underlying ancestral populations, incorporate reference individuals of known ancestry to improve ancestry estimates, and penalize small ancestry estimates to improve model parsimony and avoid model fitting problems [Alexander and Lange 2011].

Newer software packages for calculating global ancestry proportion have been constructed to take sequencing-derived genotypes with uncertainty into account, as well as to construct population relationship trees from the data. NGSadmix [Skotte, et al. 2013] is an extension of the MLE framework built to accommodate genotype likelihoods often available from low-depth next-generation sequencing (NGS) data due to the uncertainty regarding the true genotypes. Although slower than ADMIXTURE, use of the genotype likelihoods outperform the hard-called genotypes, when sequencing depth was approximately even for all individuals and average depth was at least 0.5x. Ohana [Cheng, et al. 2017] is another new method for inferring admixture in an MLE framework which is applicable both to called genotypes and to NGS data, which also estimates population relationships using a Gaussian approximation. It selects the best covariance matrix compatible with a tree, thereby estimating a tree, and provides simple algorithms and visualization tools to obtain the evolutionary trees.

There are also software packages that explicitly map spatial differences in ancestry. These methods were developed to approximate the geographic location of admixed populations. SPA [Yang, et al. 2012] is a probabilistic model for the spatial structure of genetic variation, which explicitly models how the allele frequency of each SNP changes as a function of the location of the individual in geographic space (where the allele frequency is a function of the x and y coordinates of an individual on a map). This approach detects SNPs with steep

geographic gradients in allele frequency that suggest the SNPs have been under selection. Geographic Ancestry Positioning (GAP) uses genotypes to infer local spatial distances and applies them to a global space [Bhaskar, et al. 2017]. This method has been extended to an association test, which uses an allele frequency smoothing technique using the spatial coordinates and incorporates that information to test each SNP in an inverse regression of the genotype against the trait, conditional on the estimated allele frequency.

Methods for observing and clustering ancestral groups—Several methods infer PS, such as MDS or PCA, using singular value or eigenvector decomposition. These methods use genome-wide level data to summarize genetic variance in variables that can be plotted to visualize genetic relationships between samples. When visualized alongside known reference populations, ancestral background, and therefore PS, among the sample can be identified. Other scenarios can also cause clustering of samples, including kinship and genotyping batch effects. Thus visualization of the components with reference population anchors is strongly recommended to ensure clustering by continental ancestry in reference populations is as expected. PCA is implemented in EIGENSTRAT [Price, et al. 2006], as well as MDS implemented in PLINK software [Purcell, et al. 2007]. The output from PCA and MDS is often very similar, as illustrated in Figure 1.

Recently several faster methods for computation of PCs have been developed that use randomized matrix algorithms and parallelized matrix multiplication. These methods include FlashPCA [Abraham and Inouye 2014] and FastPCA [Galinsky, et al. 2016] implemented in EIGENSOFT as fastmode option. FastPCA computational time scaled linearly with increasing sample size as opposed to other methods that have shown cubic or quadratic increases. Analysis of 100,000 individuals and 100,000 SNPs with FastPCA on a single computer required less than an hour and only 3.2GB memory, while flashPCA required nearly 10 hours and 40GB to compute across 30,000 samples. The LASER program [Wang, et al. 2014; Wang, et al. 2015] is designed to handle low-coverage sequence reads to perform PCA. When combined with genotype imputation, LASER 2.0 can accurately estimate fine-scale genetic ancestry, and is implemented on a web server (http://laser.sph.umich.edu/) [Taliun, et al. 2017].

The SNPweights method [Chen, et al. 2013] assigns weights to the individual SNPs in the analysis by population. Weights for SNPs are pre-computed in the reference panel and those weights can be applied to the sample to infer ancestry without having to gain access to the raw genotypes of the reference panel. This is similar to the approaches utilizing AIMs, however, this incorporates more SNPs which improves accuracy when genome-wide SNPs are available.

Principal Components Analysis with Related individuals (PC-AiR) [Conomos, et al. 2015] infers PS in the presence of related individuals. This method identifies an unrelated subset of individuals that represents the ancestral diversity of the sample and computes PCs in this subset and projects PCs onto the remainder of the sample. This approach does not require reference samples to be included for adequate performance, but does perform better when incorporating kinship coefficients when defined pedigree structure among samples is unknown.

Extensions of PCA have been developed to handle complex ancestral scenarios are also available. PCAmask [Moreno-Estrada, et al. 2013] and subspace PCA (ssPCA) [Johnson, et al. 2011] were developed to address the complex recent admixture of indigenous and Native American populations. These approaches analyze genomic segments consistent with a single inferred continental population (virtual genomes). The PCAmask approach extends upon ssPCA by utilizing phased haplotype data, allowing use of genomic regions that are ancestrally heterozygous.

Genomic Control—Another popular PS method is Genomic control (GC), which controls the inflation of test statistics and can also be used to detect PS. It was developed for dichotomous traits [Devlin, et al. 2004; Devlin and Roeder 1999; Devlin, et al. 2001] and then extended to quantitative traits [Bacanu, et al. 2002]. At least 100 uncorrelated SNPs should be genotyped for GC, and these SNPs should not be associated with the trait of interest. The goal of GC is to quantify the bias in the data, either due to confounding, experimental errors, cryptic relatedness, or other causes. When SNPs in the GC set are associated with the trait, then their test statistics represent the alternative hypothesis and appear biased compared to the distribution expected under the null hypothesis. Thereby, if non-null SNPs are used to calculate the GC correction, then the correction will be conservative and associations will be more difficult to detect.

GC adjusts the observed distribution of the test statistic Y for tests of association between these null markers and the trait. Under the null hypothesis of no association, the Armitage trend test for association of SNPs with traits is asymptotically equal to a chi-square distribution. When there is PS, the test statistic is inflated by a factor, λ. Therefore, the statistic (Y) results from the inflation of the Armitage trend test, which can be written as $Y = \lambda\chi_1^2$. λ is then calculated as $\hat{\lambda}$ = median(Y_1, Y_2, …, Y_L)/0.4549 or $\hat{\lambda}$ = mean(Y_1, Y_2, …, Y_L)/1 since the median and mean of χ_1^2 are 0.4549 and 1, respectively. By estimating λ from the unassociated SNPs and using Y_i/λ to calculate p-values in place of Y_i, for i markers, the effect of PS on p-values will be removed, reestablishing the χ_1^2 distribution over a large number of SNPs.

Additionally, $\hat{\lambda}$ provides a convenient quality control statistic for assessing whether association tests for GWAS data are confounded. This is done by checking if $\hat{\lambda}$ is much different from 1 in the lower 90% of ranked test statistics, where smaller p-values are excluded to avoid apparent inflation due to true associations of SNPs with traits. Large values for $\hat{\lambda}$ (values of $\hat{\lambda} < 1.05$ are considered benign, and values of 1.2 or more are tolerated for very large studies of highly polygenic traits with sample sizes of hundreds of thousands of participants) indicate that tests of association are confounded by some phenomena, which may include PS.

Additionally, other types of systematic differences between the data from groups of subjects can also cause large $\hat{\lambda}$, such as nonrandom genotyping error that might arise due to merging GWAS data from different genotyping experiments, nonrandom differences in DNA quality between study samples, or other unmeasured confounders. This procedure is implemented in PLINK and is straightforward to calculate with any statistical software [Purcell, et al. 2007].

GC is reported to be ineffective if too few loci (< 100) are used and may decrease power if too many loci (> 500) are used [Marchini, et al. 2004]. Recent GWAS studies usually use λ calculated from genome-wide SNPs as an important post-analysis diagnostic statistic, and to protect against excess type I error. There is substantial variation in estimates of λ that depend on the set of markers chosen, and this may also decrease power if PS is extreme [Kohler and Bickeboller 2006; Zhang, et al. 2008]. GC can also be conservative if AIMs are used instead of random markers [Epstein, et al. 2007].

An alternative approach, GCF, which is a modified version of GC that uses the F distribution, has been shown to be more appropriate than GC in some extreme examples of PS [Dadd, et al. 2009; Devlin, et al. 2004]. GC also does not correct effect size estimates, although it can be used to correct confidence intervals, and as a result odds ratios or linear regression coefficients will be unreliable after GC is applied, even though the test statistics and p-values have been adjusted.

In scenarios where meta-analysis is being performed across several GWAS, GC corrections can be performed within each study and then again in the meta-analysis results. This double GC correction adjusts the set of test statistics across all markers within each study by a GC inflation factor. It then calculates a combined statistic across studies at each marker, and adjusts all combined statistics across the genome by the corresponding GC inflation factor. It has been suggested that PCA correction is more effective than the double GC correction in meta-analysis [Wang, et al. 2012b]. In the case where population stratification exists, using the double GC method usually results in much lower power than using the PCA correction in meta-analysis, even when the causal SNP does not have significant allele frequency differentiation in the subpopulations.

LD Score Regression—LD Score Regression [Bulik-Sullivan, et al. 2015] is a method utilizing summary association statistics from a GWAS to determine whether inflation of the test statistics is due to a true polygenic signal or bias. LD scores are computed in a sequenced reference panel with similar LD structure as in the GWAS by calculating the strength of tagging by SNPs within a 1CM window. LD score regression can be used to estimate the mean contribution of confounding bias to the inflation in the test statistics; i.e. to indicate post-hoc whether there is residual cryptic relatedness or population stratification remaining in the dataset. However, the model assumes that there is no systematic correlation between F_{st} and LD Score, which may not be the case when there is selection. It was demonstrated that the average LD Score regression intercept was approximately equal to the λ in simulations with PS. Because λ increases with sample size in the presence of polygenicity, the gain in power obtained by correcting test statistics with the LD Score regression intercept instead of λ will become even more substantial for larger GWAS.

Subtle Stratification (PC Loading regression)—An approach for correcting residual inflation of test statistics is PC loading regression [Bhatia 2016], that integrates the concept of weighting SNPs according to their contribution to PCs (i.e. total genetic variance) and also incorporates rare variant haplotypes. These rare variants are often omitted from PCA during the LD-pruning process. The slope of this PC loading regression provides an estimate of the magnitude of PS. It has been suggested that rare haplotypes can better capture subtle

PS [Bhatia, et al. 2016], a concept which has been leveraged in several fundamental approaches in human genomics (i.e. rare variant enrichment in extended pedigrees [Browning and Thompson 2012], the continued relevance of linkage analysis [Ott, et al. 2015; Teare and Santibanez Koref 2014], haplotype length investigations for selection [Lappalainen, et al. 2010].

Cryptic relatedness and population stratification—As genetic studies have grown larger over time, so has the likelihood of recruiting related individuals or those who share extended relationships unbeknownst to the investigators. These cryptic relationships can also influence association statistics much like PS. The relatedness between two individuals is most frequently expressed in terms of the probability that they share zero, one or two alleles that are inherited identical-by-descent (IBD). However, as sample sizes have increased, construction of IBD matrices has become more computationally intensive, and determining whether to use relatedness as an exclusion process or to model it explicitly has been a subject of much debate. While PCA may detect and account for some relatedness, it may not adequately control this or may result in loss of power beyond those methods directly accounting for these relationships. Use of mixed models to account for cryptic relatedness has been one popular strategy for retaining as many samples as possible, as has reconstruction of pedigrees using IBD information. Mixed models have been shown to have better overall performance than PCA in the presence of association [Wang, et al. 2013].

PS can be thought of as a special case of cryptic relatedness, where participants who share parental ancestral populations are more closely related to each other than they are to participants who arise from different populations. In that conceptual framework for PS all participants in the study are connected by a large latent pedigree, with the ancestors that connect them unobserved. The number of meiosis that separate closely related people are small, are larger for distantly related people from the same population, and are much larger for pairs of people from distinct continental populations with long coalescence times. A group of methods that are designed to leverage this property of genetic data for quantitative traits are the linear mixed models, which can mitigate both the issues that arise when there is cryptic relatedness [Voight and Pritchard 2005] and PS in association studies in one procedure [Kang, et al. 2010; Kang, et al. 2008; Listgarten, et al. 2012; Zhou and Stephens 2012]. These approaches were originally developed for model organism studies in multiple inbred and outbred lines where many spurious results were initially observed, but were then non-significant after application of the mixed model methods. In a recent review of mixed model analyses, Martin and Eskin describe the formulation of these analyses, and show that mixed models produce smaller $\hat{\lambda}$ statistics than PCA correction or removal of related subjects for a range of quantitative traits in a structured population from Finland [Martin and Eskin 2017].

In addition to confounding genotype-phenotype associations, PS may also distort estimates of trait heritability. Heritability for a particular trait may be described as the proportion of trait variability explained by genetic variants. Though historically measured in family studies, newer methods have been developed to estimate heritability from genome-wide data in population-based studies [Lee, et al. 2011; Yang, et al. 2010; Yang, et al. 2011]. However, Dandine-Roulland *et al.* warn that model adjustment for ancestry inferred from genomic data

does not adequately correct for PS bias in population-based heritability estimates [Dandine-Roulland, et al. 2016].

Local Ancestry Methods

Local ancestry methods are used to identify ancestral origins of chromosomal regions. In two-way admixture, such as African Americans, any given genetic locus will have exactly 0%, 50% or 100% European derived alleles corresponding to 0, 1, or 2 copies. Accurate inference of local ancestry depends on number of generations since the admixture event, number of admixture events, number of ancestral populations involved across admixture events, and availability of reference data that represent the ancestral populations. Methods for local ancestry inference may be divided into two broad classes: 1) methods which do not model linkage disequilibrium and 2) methods that leverage linkage disequilibrium (LD). Key points regarding each software discussed here are summarized in Table 2.

Methods that do not model LD

Early methods for local ancestry inference included STRUCTURE/MaldSoft, ADMIXMAP, ANCESTRYMAP, and ADMIXPROGRAM are based on variations of first-level HMM where the goal is to make inferences on a series of hidden states (local ancestry) based on observable states (alleles and allele frequencies from ancestral populations) [Falush, et al. 2003; Hoggart, et al. 2004; McKeigue 1998; Montana and Hoggart 2007; Patterson, et al. 2004; Zhu, et al. 2006]. A key assumption of the HMM models are that the observed states, or alleles, are independent of each other, conditional upon the hidden states, the ancestry source for each allele. These methods rely on unlinked AIMs. These early methods are able to infer continental ancestry throughout the genome (African and European), with the resolution limited by the number of independent AIMs available and computational tractability.

The Local Ancestry in admixed Populations (LAMP) method uses sliding windows of contiguous independent SNPs to infer local ancestry[Sankararaman, et al. 2008]. It first calculates an optimal window length such that the probability that a given window has a recombination event is small and assumes that alleles in the window are derived from only one ancestry. It then uses a clustering algorithm known as Iterated Conditional Modes (ICM) on each of these windows to infer ancestry on each marker on the window, followed by a majority vote across overlapping windows to call ancestry. Advantages of this method over previous methods include: faster run times, capability of handling GWAS data, improved accuracy, ability to infer local ancestry even in the absence of ancestral reference data, and incorporation of ancestral allele-frequency data (LAMP-ANC) when it is available for even more accurate predictions. These methods are optimized to make ancestry calls for admixed populations with two-way admixture between distant ancestral populations such as Africans and Europeans. These methods are inaccurate if inferences are made on closely related populations.

WINPOP modifies and extends the LAMP method to provide inference of local ancestry not only in admixed individuals with distant ancestry, but also between closely related populations [Pasaniuc, et al. 2009]. It uses a sliding window like LAMP with two important

distinctions: 1) it adaptively determines window length for each location and 2) it allows for up to one recombination event to occur within each window. The method provides more accurate results than LAMP, and LD-based methods such as SABER and HAPAA (discussed below) across distant and closely related two-way admixtures [Sankararaman, et al. 2008; Sundquist, et al. 2008; Tang, et al. 2006]. The greatest gains in accuracy were reported to occur in closely related populations. Despite this improvement, the method has up to 91% accuracy for closely related populations.

Methods that model LD

LD based methods for local ancestry inference assume that there may be haplotypes unique to a given population. SABER is one of these methods and uses a 'Markov-switching model,' also known as Markov Hidden Markov Models (MHMM)s [Tang, et al. 2006]. Previous HMMs were incapable of handling LD between markers as modeling haplotypes within ancestral populations in the HMM framework would be computationally intractable. A similar approach, HMM-based Analysis of Polymorphisms in Admixed Ancestries (HAPAA), uses hierarchical HMMs to model LD, displays lower error rates than SABER, and also has features that evaluate the effect of genetic divergence between ancestral populations and time-to-admixture [Sundquist, et al. 2008].

HAPMIX is a haplotype-based HMM method that achieves high accuracy and has a strict assumption of two ancestral populations [Price, et al. 2009]. It utilizes the population genetic model of Li and Stephens and phased haplotypes from unadmixed ancestral populations as references to infer local ancestry [Li and Stephens 2003]. HAPMIX, like HAPAA uses HMM to explicitly model LD to make local ancestry inference with a few key differences. It allows some margin of error for miscopying ancestry segments from the wrong population. It also allows for unphased data for the admixed population and attempts to account for phase-flip errors on ancestry inference. These features along with the use of dense SNPs allows it to make inference on smaller stretches of chromosome, which is where ancient admixture is likely to be detected. However, the requirement for phased haplotypes from unadmixed ancestral populations and the specification of many parameters limits its use in less-studied populations.

LAMP-LD improves on existing methods by proposing a model of local ancestry inference that extends to multi-way admixed populations with significantly reduced error rates [Baran, et al. 2012]. Like HAPAA and HAPMIX it models the haplotype structure using HMMs, but with a fixed-size state-space. This is the only method that uses a fast-approximation of the Li and Stephens model to realize ancestral haplotype structures. Additionally, LAMP-LD estimates its parameters from the reference haplotypes rather than relying on user-specification which greatly reduces parameter misspecification. Furthermore, it combines the window-based method originally developed for LAMP with an HMM method that relaxes the no-recombination limitation, improving speed and accuracy in three-way admixed populations. Another extension to LAMP-LD is LAMP-HAP, which further leverages pedigree information from trio data to provide local ancestry estimates with greater accuracy.

RFMix departs from the HMM-extension framework discussed above to a discriminative approach to explicitly models ancestry along an admixed chromosome given known reference haplotypes or inferred ancestry [Maples, et al. 2013]. In the inference algorithm for RFMix, phased reference chromosomes are first divided into windows of equal sizes by genetic distance. A random forest is then trained within each window to classify ancestry. The random forest is then applied to the corresponding window of admixed chromosome to generate fractional votes, which are then summed to generate posterior probabilities for ancestry within each window. Posterior probabilities from consecutive windows are then put through max marginalization of the forward-backward posterior probabilities to infer the most likely sequence of ancestry across windows. The method is faster than LAMP-LD or LAMP-HAP and provides more accurate estimates of local ancestry. An important feature of this software that it is accurate even when reference data is limited. The algorithm is also able to incorporate inferred ancestry segments from the admixed chromosome to further augment the training set in an iterative process.

Admixture mapping

Admixture mapping can be used to identify disease causing loci in admixed populations. Admixture mapping is an ideal approach for studying diseases with differential prevalence across ancestral populations where the disparity is heritable. Methods for admixture mapping have been covered extensively in the following review articles [Seldin, et al. 2011; Shriner 2013; Winkler, et al. 2010]. Briefly, case-only and case-control admixture mapping strategies have been widely used in the past. While case-only admixture mapping strategies can provide greater sensitivity in detecting disease loci, they are also particularly prone to false-positive signals. Case-control admixture mapping strategies provide a stronger control of type I error. Both case-only and case-control admixture strategies have advantages over GWAS for multiple testing. Because ancestral LD blocks tend to be much longer than short-range LD, the number of independent tests is drastically reduced with admixture mapping.

In addition to the traditional admixture mapping strategies, at least two joint test frameworks leverage local ancestry to increase power for gene discovery. The first joint method is implemented in Mixscore, which combines a case-only admixture test statistic with a SNP association test into a single one-degree of freedom chi-square test [Pasaniuc, et al. 2011]. This test is more powerful for discovery than the Armitage trend test, the SNP association test while conditioning on local ancestry, case-only association test, and also the two-degree of freedom chi-square joint-test for the sum of SNP and admixture mapping association test.

Another joint test is the BMIX that uses a Bayesian framework to model posterior probabilities from admixture mapping as prior probabilities for association testing to reduce multiple testing [Shriner, et al. 2011]. In simulations the authors show BMIX to be more powerful than the Mixscore approach.

STRATEGY

A summary of most methods described in this article is presented in Table 2, with an outline of capabilities and limitations for each approach.

Investigations of genetic traits in humans are observational studies where researchers do not perform mating experiments, control the environment for the organism, or induce mutations such as in experimental studies with model organisms. As a result of the observational nature of the research, as for any epidemiologic investigation, care must be taken when planning the ascertainment of subjects and the statistical analysis of the genetic data to control for confounders.

One of the most common and important considerations regarding potential confounding in human genetic epidemiology is whether a sample of subjects under study includes persons of mixed ancestry or groups of subjects with distinct ancestral backgrounds. When there is a difference across ancestral groups in the probability of ascertaining a subject with the phenotype of interest or a difference in the distribution of a quantitative trait between ancestral groups, then any genetic variant with a difference in allele frequency across ancestral groups might seem to be associated with the trait if tests of association are carried out without accounting for ancestry [Freedman, et al. 2004]. Failure to adjust for PS properly can lead to excess false positive results or cause loss of power [Cardon and Palmer 2003; Marchini, et al. 2004]. As a result of the associations of alleles with ancestry, the degree of confounding is related to the sample size of the study, such that larger studies are more acutely affected [Marchini, et al. 2004].

The design of a genetic study can involve one of several sampling strategies and stages. The sampling approach is most likely determined by properties of the trait and the availability of existing studies with biological specimens from the study subjects. For example, when studying a trait with an onset that is typically early in life it may not be feasible to recruit large numbers of healthy unrelated control children, since the parents of healthy children are often less motivated to participate in research than parents of cases. As a result a family-based design utilizing the TDT or a related statistic may be more efficient. Conversely, for a trait with an onset late in life, other relatives in the family may not be available, and so a case-control study may be easiest to conduct. For a case-control study, an ideal sample of controls would have the same potential as the cases for exposure to risk factors, and if the controls had manifested the trait they would be selected as cases for the study. This principal is violated when there is PS in the data that is also related to the trait of interest through a difference in prevalence for the trait and alleles at many loci in the parental populations.

When designing a genetic study, some effort should be expended to identify the ancestry of subjects before genotyping commences. For example, an option is to require all members of the study to report the ethnicity of all four grandparents for eligibility and crude quantification of ancestry [Velez, et al. 2008]. However, in certain situations this may misclassify an individual’s actual ancestral background if ancestry is a cultural rather than a genetic classifier, as is sometimes the case in Hispanic populations.

One aspect of investigating traits in admixed samples is the difficulty of performing replication studies. Once an association with a particular marker has been observed, a second round of genotyping is usually performed in an independent sample of subjects to verify the signal. To account for PS using global estimates of ancestry, several dozen AIMs may be necessary. This could increase the cost of the replication study by several times,

limiting the sample size that may be investigated and consequently the chances of successful replication. When there are a small number of SNPs of interest to replicate, we advocate using local ancestry estimates from the candidate marker and several nearby flanking AIMs to adjust for PS. This issue arises in consortium studies of GWAS data in admixed populations, and can be a challenge to coordinate replication efforts using global estimates of ancestry in previously ungenotyped subjects. Alternatively if participants from non-admixed parental populations are available, they may be analyzed without adjustment for PS [Franceschini, et al. 2013; Monda, et al. 2013].

Another consideration is whether the plan for genotyping accommodates the idiosyncrasies of the study sample. If the study design is investigating candidate genes, then a panel of approximately 30 AIMs may be necessary to quantify global ancestry in African Americans, and more in populations with more complex demographic histories. Alternately, nearby AIMs not in LD with the gene regions may be added to the genotyping panel to call local ancestry with a method such as LAMP or HAPMIX. Both of these designs require that suitable reference panels of genotypes are available from the appropriate ancestral populations. If this is the case, then an agnostic method such as STRUCTURE might be used, with a scan through possible values for the number of ancestral subpopulations. If GWAS data are being generated, then MDS or PCA can be applied to summarize continuous axes of ancestral variation and adjust for confounding by PS.

Proper use of these methods requires a working understanding of population genetics principles and association statistics for genetic epidemiology. One of the most important considerations; however, is the study design employed and how that design will work in concert with the analytic methods to produce reliable results.

COMMENTARY

In this article we focused on PS methods and their applications in human disease mapping. In addition, many of the methods we present here are also used in experimental populations of animals and plants, agriculture, and ecology [Bomblies, et al. 2010; Galvan, et al. 2011].

PS is an extensively studied area of research. Other than the general PS methods mentioned previously, some PS methods are designed for some special situations. For example, some methods have the ability to conduct association tests for a combination of pedigree and unrelated samples while correcting for PS [Chung, et al. 2010; Thornton and McPeek 2010; Zhu, et al. 2008]. There are also several early methods that used coarse sets of genetic markers that specifically target admixed populations [Hoggart, et al. 2003; Montana and Pritchard 2004; Patterson, et al. 2004]. Interested readers in admixture mapping and population stratification in general can consult other recent reviews on disease mapping in admixed populations [Astle and Balding 2009; Price, et al. 2010; Seldin, et al. 2011].

Genetic research is expanding into more diverse populations, and PS will continue to be important in human genetic studies. It is also becoming clear that the rare alleles carried by each population are unique, and traits may have distinct etiologies in various human populations [Gravel, et al. 2011]. This may be the cause for the apparent failures of some

previously observed associations to replicate when the associated SNPs are assayed in other populations. Other causes that are related to the differences between populations are also likely to cause apparent failure to replicate at specific SNPs, such as differences in LD, environmental exposures, and different frequencies of genetic modifiers.

References

Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. PLoS One. 2014; 9(4):e93766. [PubMed: 24718290]

Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011; 12:246. [PubMed: 21682921]

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19(9):1655–64. [PubMed: 19648217]

Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. Statistical Science. 2009; 24(4):451–471.

Bacanu SA, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. Genet Epidemiol. 2002; 22(1):78–93. [PubMed: 11754475]

Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguez-Cintron W, Chapela R, Ford JG, Avila PC, et al. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics (Oxford, England). 2012; 28:1359–1367.

Bayless TM, Brown E, Paige DM. Lactase Non-persistence and Lactose Intolerance. Curr Gastroenterol Rep. 2017; 19(5):23. [PubMed: 28421381]

Bell GI, Horita S, Karam JH. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. Diabetes. 1984; 33(2):176–83. [PubMed: 6363172]

Benn-Torres J, Bonilla C, Robbins CM, Waterman L, Moses TY, Hernandez W, Santos ER, Bennett F, Aiken W, Tullock T, et al. Admixture and population stratification in African Caribbean populations. Ann Hum Genet. 2008; 72(Pt 1):90–8. [PubMed: 17908263]

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004; 74(6):1111–20. [PubMed: 15114531]

Bhaskar A, Javanmard A, Courtade TA, Tse D. Novel probabilistic models of spatial genetic ancestry with applications to stratification correction in genome-wide association studies. Bioinformatics. 2017; 33(6):879–885. [PubMed: 28025204]

Bhatia G, Furlotte NA, Loh P-R, Liu X, Finucane HK, Gusev A, Price A. Correcting subtle stratification in summary association statistics. bioRxiv. 2016:076133.

Bhatia G, Gusev A, Loh P-R, Finucane HK, Vilhjalmsson BJ, Ripke S, Purcell S, Stahl E, Daly M, de Candia TR, et al. Subtle stratification confounds estimates of heritability from rare variants. bioRxiv. 2016

Bomblies K, Yant L, Laitinen RA, Kim ST, Hollister JD, Warthmann N, Fitz J, Weigel D. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. PLoS Genet. 2010; 6(3):e1000890. [PubMed: 20361058]

Browning SR, Thompson EA. Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics. 2012; 190(4):1521–31. [PubMed: 22267498]

Brucato N, Tortevoye P, Plancoulaine S, Guitard E, Sanchez-Mazas A, Larrouy G, Gessain A, Dugoujon JM. The genetic diversity of three peculiar populations descending from the slave trade: Gm study of Noir Marron from French Guiana. C R Biol. 2009; 332(10):917–26. [PubMed: 19819412]

Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A. 2010; 107(2):786–91. [PubMed: 20080753]

Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet. 2015; 96(1):37–53. [PubMed: 25529636]

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015; 47(3):291–5. [PubMed: 25642630]

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. Demonstrating stratification in a European American population. Nat Genet. 2005; 37(8):868–72. [PubMed: 16041375]

Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet. 2003; 361(9357):598–604. [PubMed: 12598158]

Carmi S, Hui KY, Kochav E, Liu X, Xue J, Grady F, Guha S, Upadhyay K, Ben-Avraham D, Mukherjee S, et al. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat Commun. 2014; 5:4835. [PubMed: 25203624]

Chen CY, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved ancestry inference using weights from external reference panels. Bioinformatics. 2013; 29(11):1399–406. [PubMed: 23539302]

Cheng YJ, Mailund T, Nielsen R. Fast admixture analysis and population tree estimation for SNP and NGS data. Bioinformatics. 2017

Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai HJ, Naqvi M, Phong A, Ung N, et al. Population stratification confounds genetic association studies among Latinos. Hum Genet. 2006; 118(5):652–64. [PubMed: 16283388]

Chung RH, Schmidt MA, Morris RW, Martin ER. CAPL: a novel association test using case-control and family data and accounting for population stratification. Genet Epidemiol. 2010; 34(7):747–55. [PubMed: 20878716]

Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet Epidemiol. 2015; 39(4):276–93. [PubMed: 25810074]

Dadd T, Weale ME, Lewis CM. A critical evaluation of genomic control methods for genetic association studies. Genet Epidemiol. 2009; 33(4):290–8. [PubMed: 19051284]

Dandine-Roulland C, Bellenguez C, Debette S, Amouyel P, Genin E, Perdry H. Accuracy of heritability estimations in presence of hidden population stratification. Sci Rep. 2016; 6:26471. [PubMed: 27220488]

Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, Francioli LC, Hottenga JJ, Karssen LC, Estrada K, et al. Improved imputation quality of low-frequency and rare variants in European samples using the ‘Genome of The Netherlands’. Eur J Hum Genet. 2014; 22(11):1321–6. [PubMed: 24896149]

Devlin B, Bacanu SA, Roeder K. Genomic Control to the extreme. Nat Genet. 2004; 36(11):1129–30. author reply 1131. [PubMed: 15514657]

Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55(4):997–1004. [PubMed: 11315092]

Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol. 2001; 60(3):155–66. [PubMed: 11855950]

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. Identification of a variant associated with adult-type hypolactasia. Nat Genet. 2002; 30(2):233–7. [PubMed: 11788828]

Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. Am J Hum Genet. 2007; 80(5):921–30. [PubMed: 17436246]

Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, Saxonov S, Avey L, Wojcicki A, Pe’er I, Mountain J. Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet. 2010; 6(6):e1000993. [PubMed: 20585627]

Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003; 164(4):1567–87. [PubMed: 12930761]

Franceschini N, Fox E, Zhang Z, Edwards TL, Nalls MA, Sung YJ, Tayo BO, Sun YV, Gottesman O, Adeyemo A, et al. Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. Am J Hum Genet. 2013; 93(3):545–54. [PubMed: 23972371]

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004; 36(4):388–93. [PubMed: 15052270]

Friedrich DC, Santos SE, Ribeiro-dos-Santos AK, Hutz MH. Several different lactase persistence associated alleles and high diversity of the lactase gene in the admixed Brazilian population. PLoS One. 2012; 7(9):e46520. [PubMed: 23029545]

Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, Price AL. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. Am J Hum Genet. 2016; 98(3):456–72. [PubMed: 26924531]

Galvan A, Vorraro F, Cabrera W, Ribeiro OG, Starobinas N, Jensen JR, dos Santos Carneiro P, De Franco M, Gao X, Ibanez OC, et al. Association study by genetic clustering detects multiple inflammatory response loci in non-inbred mice. Genes Immun. 2011; 12(5):390–4. [PubMed: 21346777]

Gao X, Martin ER. Using allele sharing distance for detecting human population stratification. Hum Hered. 2009; 68(3):182–91. [PubMed: 19521100]

Gao X, Starmer J. Human population structure detection via multilocus genotype clustering. BMC Genet. 2007; 8:34. [PubMed: 17592628]

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. Genomes Project C. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. [PubMed: 23128226]

Goddard KA, Hopkins PJ, Hall JM, Witte JS. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet. 2000; 66(1):216–34. [PubMed: 10631153]

Gomes KF, Santos AS, Semzezem C, Correia MR, Brito LA, Ruiz MO, Fukui RT, Matioli SR, Passos-Bueno MR, Silva ME. The influence of population stratification on genetic markers associated with type 1 diabetes. Sci Rep. 2017; 7:43513. [PubMed: 28262800]

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Genomes P, Bustamante CD. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A. 2011; 108(29):11983–8. [PubMed: 21730125]

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al. A draft sequence of the Neandertal genome. Science. 2010; 328(5979):710–22. [PubMed: 20448178]

Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 2013; 9(6):e1003521. [PubMed: 23754952]

Higasa K, Miyake N, Yoshimura J, Okamura K, Niihori T, Saitsu H, Doi K, Shimizu M, Nakabayashi K, Aoki Y, et al. Human genetic variation database, a reference database of genetic variations in the Japanese population. J Hum Genet. 2016; 61(6):547–53. [PubMed: 26911352]

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM. Control of confounding of genetic associations in stratified populations. Am J Hum Genet. 2003; 72(6):1492–1504. [PubMed: 12817591]

Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. Am J Hum Genet. 2004; 74(5):965–78. [PubMed: 15088268]

Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet. 2009; 10(9):639–50. [PubMed: 19687804]

Homa DM, Mannino DM, Lara M. Asthma mortality in U.S. Hispanics of Mexican, Puerto Rican, and Cuban heritage, 1990–1995. Am J Respir Crit Care Med. 2000; 161(2 Pt 1):504–9. [PubMed: 10673193]

Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, Danecek P, Malerba G, Trabetti E, Zheng HF, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun. 2015; 6:8111. [PubMed: 26368830]

Ingram CJ, Elamin MF, Mulcare CA, Weale ME, Tarekegn A, Raga TO, Bekele E, Elamin FM, Thomas MG, Bradman N, et al. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? Hum Genet. 2007; 120(6):779–88. [PubMed: 17120047]

Ingram CJ, Raga TO, Tarekegn A, Browning SL, Elamin MF, Bekele E, Thomas MG, Weale ME, Bradman N, Swallow DM. Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. J Mol Evol. 2009; 69(6): 579–88. [PubMed: 19937006]

Jenkins DL, Davis LG, Stafford TW Jr, Campos PF, Hockett B, Jones GT, Cummings LS, Yost C, Connolly TJ, Yohe RM 2nd, et al. Clovis age Western Stemmed projectile points and human coprolites at the Paisley Caves. Science. 2012; 337(6091):223–8. [PubMed: 22798611]

Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ, Tang H. Ancestral components of admixed genomes in a Mexican cohort. PLoS Genet. 2011; 7(12):e1002410. [PubMed: 22194699]

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42(4):348–54. [PubMed: 20208533]

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008; 178(3):1709–23. [PubMed: 18385116]

Kawai Y, Mimori T, Kojima K, Nariai N, Danjoh I, Saito R, Yasuda J, Yamamoto M, Nagasaki M. Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals. J Hum Genet. 2015; 60(10):581–7. [PubMed: 26108142]

Kim K, Bang SY, Lee HS, Bae SC. Construction and application of a Korean reference panel for imputing classical alleles and amino acids of human leukocyte antigen genes. PLoS One. 2014; 9(11):e112546. [PubMed: 25398076]

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet. 1988; 43(4):520–6. [PubMed: 3177389]

Kodaman N, Aldrich MC, Smith JR, Signorello LB, Bradley K, Breyer J, Cohen SS, Long J, Cai Q, Giles J, et al. A small number of candidate gene SNPs reveal continental ancestry in African Americans. Ann Hum Genet. 2013; 77(1):56–66. [PubMed: 23278390]

Kohler K, Bickeboller H. Case-control association tests correcting for population stratification. Ann Hum Genet. 2006; 70(Pt 1):98–115. [PubMed: 16441260]

Lappalainen T, Salmela E, Andersen PM, Dahlman-Wright K, Sistonen P, Savontaus ML, Schreiber S, Lahermo P, Kere J. Genomic landscape of positive natural selection in Northern European populations. Eur J Hum Genet. 2010; 18(4):471–8. [PubMed: 19844263]

Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011; 88(3):294–305. [PubMed: 21376301]

Lewinsky RH, Jensen TG, Moller J, Stensballe A, Olsen J, Troelsen JT. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. Hum Mol Genet. 2005; 14(24):3945–53. [PubMed: 16301215]

Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; 475(7357):493–6. [PubMed: 21753753]

Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003; 165:2213–2233. [PubMed: 14704198]

Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. Improved linear mixed models for genome-wide association studies. Nat Methods. 2012; 9(6):525–6. [PubMed: 22669648]

Low-Kam C, Rhainds D, Lo KS, Provost S, Mongrain I, Dubois A, Perreault S, Robinson JF, Hegele RA, Dube MP, et al. Whole-genome sequencing in French Canadians from Quebec. Hum Genet. 2016; 135(11):1213–1221. [PubMed: 27376640]

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016; 538(7624):201–206. [PubMed: 27654912]

Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. American Journal of Human Genetics. 2013; 93:278–288. [PubMed: 23910464]

Marcheco-Teruel B, Parra EJ, Fuentes-Smith E, Salas A, Buttenschon HN, Demontis D, Torres-Espanol M, Marin-Padron LC, Gomez-Cabezas EJ, Alvarez-Iglesias V, et al. Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers. PLoS Genet. 2014; 10(7):e1004488. [PubMed: 25058410]

Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004; 36(5):512–7. [PubMed: 15052271]

Martin, LS., Eskin, E. Review: Population Structure in Genetic Studies: Confounding Factors and Mixed Models. bioRxiv. 2017. https://doi.org/10.1101/092106

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016; 48(10):1279–83. [PubMed: 27548312]

McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature. 2005; 433(7027):733–6. [PubMed: 15716951]

McKeigue PM. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. American Journal of Human Genetics. 1998; 63:241–251. [PubMed: 9634509]

Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, Martinez-Fuentes A, Comas D. Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol. 2008; 8:213. [PubMed: 18644108]

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012; 338(6104):222–6. [PubMed: 22936568]

Miclaus K, Wolfinger R, Czika W. SNP selection and multidimensional scaling to quantify population structure. Genet Epidemiol. 2009; 33(6):488–96. [PubMed: 19194989]

Monda KL, Chen GK, Taylor KC, Palmer C, Edwards TL, Lange LA, Ng MC, Adeyemo AA, Allison MA, Bielak LF, et al. A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. Nat Genet. 2013; 45(6):690–6. [PubMed: 23583978]

Montana G, Hoggart C. Statistical software for gene mapping by admixture linkage disequilibrium. Brief Bioinform. 2007; 8(6):393–5. [PubMed: 17640923]

Montana G, Pritchard JK. Statistical tests for admixture mapping with case-control and cases-only data. Am J Hum Genet. 2004; 75(5):771–89. [PubMed: 15386213]

Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martinez RJ, Hedges DJ, Morris RW, et al. Reconstructing the population genetic history of the Caribbean. PLoS Genet. 2013; 9(11):e1003925. [PubMed: 24244192]

Mulcare CA, Weale ME, Jones AL, Connell B, Zeitlyn D, Tarekegn A, Swallow DM, Bradman N, Thomas MG. The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. Am J Hum Genet. 2004; 74(6):1102–10. [PubMed: 15106124]

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017; 541(7637):302–310. [PubMed: 28102248]

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. Genes mirror geography within Europe. Nature. 2008; 456(7218):98–101. [PubMed: 18758442]

Olds LC, Sibley E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum Mol Genet. 2003; 12(18):2333–40. [PubMed: 12915462]

Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet. 2015; 16(5):275–84. [PubMed: 25824869]

Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. Bioinformatics (Oxford, England). 2009; 25:i213–221.

Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Kao WHL, Ruczinski I, Fornage M, Siscovick DS, Zhu X, et al. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS genetics. 2011; 7:e1001371. [PubMed: 21541012]

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, et al. Methods for high-density admixture mapping of disease genes. Am J Hum Genet. 2004; 74(5):979–1000. [PubMed: 15088269]

Pena SD, Di Pietro G, Fuchshuber-Moraes M, Genro JP, Hutz MH, de Kehdy FS, Kohlrausch F, Magno LA, Montenegro RC, Moraes MO, et al. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. PLoS One. 2011; 6(2):e17063. [PubMed: 21359226]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8): 904–9. [PubMed: 16862161]

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS genetics. 2009; 5:e1000519. [PubMed: 19543370]

Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010; 11(7):459–63. [PubMed: 20548291]

Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. Theor Popul Biol. 2001; 60(3):227–37. [PubMed: 11855957]

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000a; 155(2):945–59. [PubMed: 10835412]

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000b; 67(1):170–81. [PubMed: 10827107]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–75. [PubMed: 17701901]

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. Science. 2002; 298(5602):2381–5. [PubMed: 12493913]

Ruiz-Narvaez EA, Rosenberg L, Wise LA, Reich D, Palmer JR. Validation of a small set of ancestral informative markers for control of population admixture in African Americans. Am J Epidemiol. 2011; 173(5):587–92. [PubMed: 21262910]

Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. American Journal of Human Genetics. 2008; 82:290–303. [PubMed: 18252211]

Schanfield MS, Kirk RL. Further studies on the immunoglobulin allotypes (Gm, Am and Km) in India. Acta Anthropogenet. 1981; 5(1):1–21. [PubMed: 7236354]

Schlebusch CM, Skoglund P, Sjodin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MG, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012; 338(6105):374–9. [PubMed: 22997136]

Seldin MF, Pasaniuc B, Price AL. New approaches to disease mapping in admixed populations. Nature Reviews Genetics. 2011; 12:523–528.

Shriner D. Overview of admixture mapping. Current Protocols in Human Genetics. 2013; Chapter 1(Unit 1.23)

Shriner D, Adeyemo A, Rotimi CN. Joint ancestry and association testing in admixed individuals. PLoS computational biology. 2011; 7:e1002325. [PubMed: 22216000]

Silva ME, Mory D, Davini E. Genetic and humoral autoimmunity markers of type 1 diabetes: from theory to practice. Arq Bras Endocrinol Metabol. 2008; 52(2):166–80. [PubMed: 18438527]

Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. Genetics. 2013; 195(3):693–702. [PubMed: 24026093]

Spielman RS, Baur MP, Clerget-Darpoux F. Genetic analysis of IDDM: summary of GAW5 IDDM results. Genet Epidemiol. 1989; 6(1):43–58. [PubMed: 2659430]

Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993; 52(3):506–16. [PubMed: 8447318]

Steele CD, Court DS, Balding DJ. Worldwide F(ST) estimates relative to five continental-scale populations. Ann Hum Genet. 2014; 78(6):468–77. [PubMed: 26460400]

Sundquist A, Fratkin E, Do CB, Batzoglou S. Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Research. 2008; 18:676–682. [PubMed: 18353807]

Taliun D, Chothani SP, Schonherr S, Forer L, Boehnke M, Abecasis GR, Wang C. LASER server: ancestry tracing with genotypes or sequence reads. Bioinformatics. 2017

Tang D, Anderson D, Francis RW, Syn G, Jamieson SE, Lassmann T, Blackwell JM. Reference genotype and exome data from an Australian Aboriginal population for health-based research. Sci Data. 2016; 3:160023. [PubMed: 27070114]

Tang H, Coram M, Wang P, Zhu X, Risch N. Reconstructing genetic ancestry blocks in admixed individuals. Am J Hum Genet. 2006; 79(1):1–12. [PubMed: 16773560]

Teare MD, Santibanez Koref MF. Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing. Brief Funct Genomics. 2014; 13(5):378–83. [PubMed: 25024279]

Thareja G, John SE, Hebbar P, Behbehani K, Thanaraj TA, Alsmadi O. Sequence and analysis of a whole genome from Kuwaiti population subgroup of Persian ancestry. BMC Genomics. 2015; 16:92. [PubMed: 25765185]

Thomson G, Valdes AM, Noble JA, Kockum I, Grote MN, Najman J, Erlich HA, Cucca F, Pugliese A, Steenkiste A, et al. Relative predispositional effects of HLA class II DRB1-DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis. Tissue Antigens. 2007; 70(2):110–27. [PubMed: 17610416]

Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet. 2010; 86(2):172–84. [PubMed: 20137780]

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, et al. The genetic structure and history of Africans and African Americans. Science. 2009; 324(5930):1035–44. [PubMed: 19407144]

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007; 39(1):31–40. [PubMed: 17159977]

Tung JY, Do CB, Hinds DA, Kiefer AK, Macpherson JM, Chowdry AB, Francke U, Naughton BT, Mountain JL, Wojcicki A, et al. Efficient replication of over 180 genetic associations with self-reported medical data. PLoS One. 2011; 6(8):e23473. [PubMed: 21858135]

Velez DR, Fortunato SJ, Thorsen P, Lombardi SJ, Williams SM, Menon R. Preterm birth in Caucasians is associated with coagulation and inflammation pathway gene variants. PLoS One. 2008; 3(9):e3283. [PubMed: 18818748]

Vernot B, Akey JM. Complex history of admixture between modern humans and Neandertals. Am J Hum Genet. 2015; 96(3):448–53. [PubMed: 25683119]

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. African populations and the evolution of human mitochondrial DNA. Science. 1991; 253(5027):1503–7. [PubMed: 1840702]

Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 2005; 1(3):e32. [PubMed: 16151517]

Wahlund S. Composition of populations from the perspective of the theory of heredity. Hereditas. 1928; 11:65–105.

Wang C, Zhan X, Bragg-Gresham J, Kang HM, Stambolian D, Chew EY, Branham KE, Heckenlively J, Study F, Fulton R, et al. Ancestry estimation and control of population stratification for sequence-based association studies. Nat Genet. 2014; 46(4):409–15. [PubMed: 24633160]

Wang C, Zhan X, Liang L, Abecasis GR, Lin X. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am J Hum Genet. 2015; 96(6):926–37. [PubMed: 26027497]

Wang C, Zollner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet. 2012a; 8(8):e1002886. [PubMed: 22927824]

Wang K, Hu X, Peng Y. An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. Hum Hered. 2013; 76(1):1–9. [PubMed: 23921716]

Wang S, Chen W, Chen X, Hu F, Archer KJ, Liu HN, Sun S, Gao G. Double genomic control is not effective to correct for population stratification in meta-analysis for genome-wide association studies. Front Genet. 2012b; 3:300. [PubMed: 23269928]

Wang X, Zhu X, Qin H, Cooper RS, Ewens WJ, Li C, Li M. Adjustment for local ancestry in genetic association analysis of admixed populations. Bioinformatics. 2011; 27(5):670–7. [PubMed: 21169375]

Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984; 38(6):1358–1370. [PubMed: 28563791]

Weir BS, Hill WG. Estimating F-statistics. Annu Rev Genet. 2002; 36:721–50. [PubMed: 12359738]

White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, Howell FC. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature. 2003; 423(6941):742–7. [PubMed: 12802332]

Williams RC, Steinberg AG, Gershowitz H, Bennett PH, Knowler WC, Pettitt DJ, Butler W, Baird R, Dowda-Rea L, Burch TA, et al. GM allotypes in Native Americans: evidence for three distinct migrations across the Bering land bridge. Am J Phys Anthropol. 1985; 66(1):1–19. [PubMed: 3976868]

Winkler CA, Nelson GW, Smith MW. Admixture mapping comes of age. Annual Review of Genomics and Human Genetics. 2010; 11:65–89.

Wong LP, Lai JK, Saw WY, Ong RT, Cheng AY, Pillai NE, Liu X, Xu W, Chen P, Foo JN, et al. Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing. PLoS Genet. 2014; 10(5):e1004377. [PubMed: 24832686]

Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, Lam KK, Pillai NE, Sim KS, Xu H, et al. Deep whole-genome sequencing of 100 southeast Asian Malays. Am J Hum Genet. 2013; 92(1):52–66. [PubMed: 23290073]

Wright S. Systems of Mating. V. General Considerations. Genetics. 1921; 6(2):167–78. [PubMed: 17245962]

Yang BZ, Zhao H, Kranzler HR, Gelernter J. Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE. Genet Epidemiol. 2005; 28(4):302–12. [PubMed: 15782414]

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42(7):565–9. [PubMed: 20562875]

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88(1):76–82. [PubMed: 21167468]

Yang WY, Novembre J, Eskin E, Halperin E. A model-based approach for analysis of spatial structure in genetic data. Nat Genet. 2012; 44(6):725–31. [PubMed: 22610118]

Zhang F, Wang Y, Deng HW. Comparison of population-based association study methods correcting for population stratification. PLoS One. 2008; 3(10):e3392. [PubMed: 18852890]

Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44(7):821–4. [PubMed: 22706312]

Zhu X, Li S, Cooper RS, Elston RC. A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet. 2008; 82(2):352–65. [PubMed: 18252216]

Zhu X, Zhang S, Tang H, Cooper R. A classical likelihood based approach for admixture mapping using EM algorithm. Hum Genet. 2006; 120(3):431–45. [PubMed: 16896924]

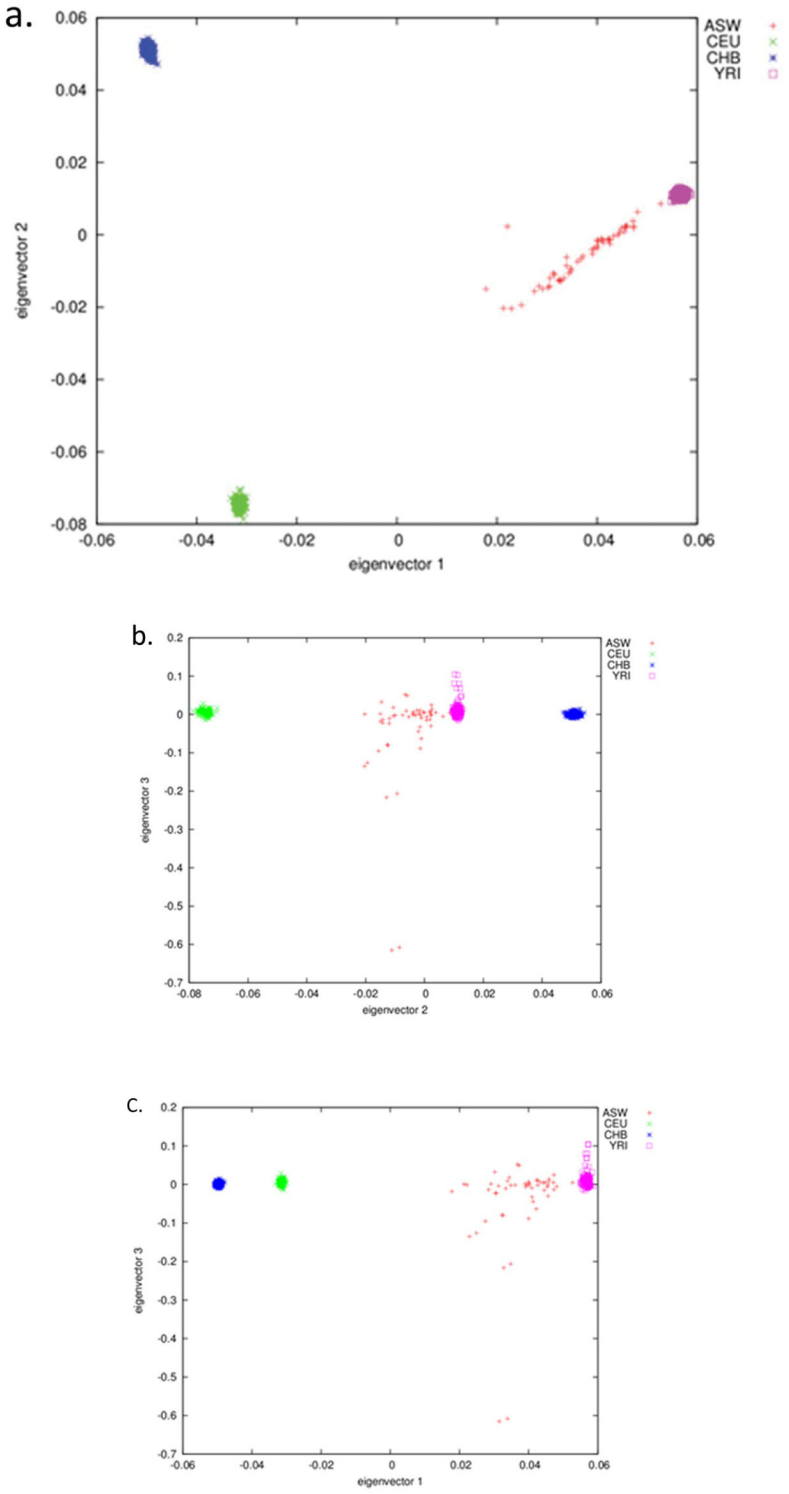

Figure 1. Principal components analysis

These figures show the clustering results using principal components analysis implemented by the Eigensoft v3.0 software with 142,616 genome-wide random autosomal SNP loci from the HapMap project (Phase 3, release 3). Only the first three eigenvectors are shown.

Note: CEU, Utah residents with ancestry from northern and western Europe; YRI, Yoruba in Ibadan, Nigeria (West Africa); CHB: Han Chinese in Beijing, China; ASW: African ancestry in Southwest USA. ASW is an admixed population.

Table 1

A numeric example of a false positive association due to population stratification.

		Phenotype			
Population	**Allele**	**Case**	**Control**	**Total**	**Association**
1	A	**270**	30	**300**	no
	B	630	70	700	
	Total	**900**	100	1000	
2	A	**80**	720	**800**	no
	B	20	180	200	
	Total	**100**	900	1000	
Pooled	A	**350**	750	**1100**	Yes
	B	650	250	900	P < .0001
	Total	**1000**	1000	2000	

Table 2

Local ancestry inference software

Program	Framework	Models LD/ Haplotype	Ancestral Population Input Format	Three-way admixture	Closely Related Populations*	Link for download
MaLDsoft/STRUCTURE	HMM-MCMC	No	-	Yes, but higher error rates	No	https://web.stanford.edu/group/pritchardlab/structure.html
ADMIXMAP	HMM-MCMC	No	Ancestral Allele Frequencies	Yes, but higher error rates	No	http://homepages.ed.ac.uk/pmckeigu/admixmap/
ANCESTRYMAP	HMM-MCMC	No	Ancestral Allele Frequencies	Two-way admixture	No	https://reich.hms.harvard.edu/software
ADMIXPROGRAM	HMM-ML	No	Ancestral Allele Frequencies	Two-way admixture	No	Available on request from Authors
SABER	MHMM	Yes, diplotyes	Phased Reference Populations	Two-way admixture	No	http://med.stanford.edu/tanglab/software/saber.html
HAPAA	Heirarchichal HMM	Yes	Phased Reference Populations	Two-way admixture	No	http://hapaa.stanford.edu
HAPMIX	MHMM-MCMC	Yes	Phased Reference Populations	Two-way admixture	No	https://reich.hms.harvard.edu/software
LAMP/LAMP-ANC	Sliding Window, ICM	No	Not Required/Anestral Allele Frequencies	Two-way admixture	No	http://lamp.icsi.berkeley.edu/lamp/
WINPOP	Adaptive Sliding Window	No	Ancestral Allele Frequencies	Two-way admixture	Yes	http://bogdan.bioinformatics.ucla.edu/software/lamp/
LAMP-LD/LAMP-HAP	Window + HMM	Yes	Phased Reference Populations	Three-way admixture	Yes	http://bogdan.bioinformatics.ucla.edu/software/lamp/
RFMix	Random Forest	Yes	Phased Reference Populations	Three-way admixture	Yes	https://sites.google.com/site/rfmixlocalancestryinference/

HMM = Hidden Markov Model; MCMC = Markov Chain Monte Carlo; ML = Maximum Likelihood; MHMM = Markov Hidden Markov Model; ICM = Iterated Conditional Mode;

*Infers ancestry accurately for closely related populations such as CHB-JPT

Table 3

Global ancestry methods and software

Program	Method	Function	Link for download
Eigensoft	PCA	Calculate PCA from genotype data	https://reich.hms.harvard.edu/software
LASER	PCA	Calculate PCA from sequencing data (low pass)	http://laser.sph.umich.edu/
FlashPCA	PCA	Rapid calculation of PCA	https://github.com/gabraham/flashpca
PC-AiR	PCA	PCA in samples that may contain cryptically related participants	http://bioconductor.org/packages/release/bioc/html/GENESIS.html
PCAmask	PCA	PCA in highly structured populations	https://github.com/armartin/ancestry_pipeline
PLINK	MDS	Calculation of multidimensional scaling variables from IBD distance matrix	http://zzz.bwh.harvard.edu/plink/
EMMA	Mixed model	Perform linear mixed model analysis for quantitative traits	http://mouse.cs.ucla.edu/emma/
GEMMA	Mixed Model	Perform linear mixed model analysis for quantitative traits	http://www.xzlab.org/software.html
EMMAX	Mixed Model	Perform linear mixed model analysis for quantitative traits more quickly than EMMA	http://genetics.cs.ucla.edu/emmax/
LD score regression	LD score regression	Calculate genomic inflation parameters accounting for LD	https://github.com/bulik/ldsc
PC loading regression	PC loading regression	Improved PS control compared with PCA	Not yet available
GAP, SCGAP	Geographic Ancestry Positioning	probabilistic spatial genetic model and ancestry localization algorithm, as well as the related population stratification correction procedure for genome-wide	https://github.com/anand-bhaskar/gap

Program	Method	Function	Link for download
		association studies, **SCGAP**,	
SNPweights	SNPweights	inferring genome-wide genetic ancestry using SNP weights precomputed from large external reference panels	https://www.hsph.harvard.edu/alkes-price/software/
NGSadmix	NGSadmix	Infer admixture proportions from NGS data	http://www.popgen.dk/software/index.php/NgsAdmix